

# Hidden Markov Model Applications in Change-Point Analysis

T. M. Luong, V. Perduca, and G. Nuel  
MAP5 Laboratory, Paris Descartes University, Paris, France

## 1 Introduction

The detection of change-points in heterogeneous sequences is a statistical challenge with many applications in fields such as finance, reliability, signal analysis, neurosciences and biology. A wide variety of literature exists for finding an ideal set of change-points for characterizing the data, in terms of both the number of change-points in a given sequence and their corresponding locations.

A conventional expression of the change-point problem is as follows: given a dataset  $X_{1:n} = (X_1, X_2, \dots, X_n)$  of real-valued observations, to find an ideal set of  $K$  non-overlapping intervals where the observations are homogeneous within each. For  $K$  segments, the change-point model expresses the distribution of  $X$  given a segmentation  $S_{1:n} = (S_1, \dots, S_n) \in \mathcal{M}_K$  as:

$$\mathbb{P}_\theta(X_{1:n}|S_{1:n}) = \prod_{i=1}^n \beta_{S_i}(X_i) = \prod_{s=1}^K \prod_{i, S_i=s} \beta_s(X_i) \quad (1)$$

where  $\beta_s$  is the emission distribution of the observed data in segment  $s$ ,  $\theta = (\theta_1, \dots, \theta_K)$  is the set of model parameters,  $S_i$  is the segment index at position  $i$ , and  $\mathcal{M}_K$  is the set of all possible combinations of  $S$  for fixed  $K \geq 2$  number of segments (for example,  $S_{1:7} = 1122233$  corresponds to  $n = 7$  observations divided into  $K = 3$  segments with two change-points: the first one between positions 2 and 3, and the second one between positions 5 and 6). We refer to this approach to the change-point model as *segment-based*, later we discuss another common approach to change-point detection that we refer to as *level-based*. For simplicity, denote  $\mathbb{P}_\theta(-)$  as  $\mathbb{P}(-)$  when the set of parameters is clear according to the context.

### 1.1 Current HMM algorithms in change-point analysis

The hidden Markov model [HMM, see Rabiner, 1989] is a commonly used tool for inference in change-point analysis. While HMMs are a conventional feature in change-point methods in bioinformatics [Fridlyand et al., 2004], it is also widely used in diverse research areas such as speech recognition [Rabiner, 1989], facial recognition [Nefian and Hayes III, 1998], financial time series [Ge and Smyth, 2000], inflation models [Chopin and Pelgrin, 2004], music classification [Kimber and Wilcox, 1997], brain imaging [Zhang et al., 2001], climate research [Hughes et al., 1999], and network security [Cho and Park, 2003].

Change-point analysis can be seen as a HMM where the data are the observations and the unknown segmentation the hidden states. HMM adaptations can therefore identify change-points by observations where a switch in hidden states is most likely to occur. A convenient feature of the HMM approaches is in inferential procedures, such as estimating the posterior marginal state

distribution  $\mathbb{P}(S|X)$ . An efficient computation in linear time of this quantity uses classical forward-backward recursions [Durbin et al., 1998]. The HMM estimation in mixture and change-point problems can be accomplished through the expectation-maximization (E-M) algorithm [Dempster et al., 1977, Bilmes, 1998], and MCMC methods, including reversible jump MCMC [Green, 1995], Gibbs sampling [Chib, 1998], and recursive algorithms [Scott, 2002]. Other approaches which use sampling to characterize the uncertainty in a HMM given the data include particle filtering [Fearnhead and Clifford, 2003] and MCMC [Guha et al., 2008]. A summary of the inferential procedures involved in HMM estimation can be found in Cappé et al. [2005]. In addition to the algorithm of linear complexity presented in this chapter, two other exact algorithms for estimating posterior distributions include a frequentist [Guédon, 2007] and Bayesian [Rigaill et al., 2011] approach which use modified versions of the forward-backward algorithm.

Many schemes for estimating the number of hidden states in the HMM have also been investigated, which in general include several penalization criteria. These include a modified Bayes Information Criterion [Zhang and Siegmund, 2007] to adjust for the number of states in previously fitted HMM as well as adaptive methods [Lavielle, 2005, Picard et al., 2005] for estimating the location and number of change-points.

Section 2 presents two different frameworks for applying HMM to change-point models, Section 3 provides a summary of two procedures for inference in change-point analysis, Section 4 provides two examples of the HMM methods on available data sets, Section 5 provides a short summary of HMM and other change-point methods for current genomics studies, and Section 6 provides a short conclusion and discussion.

## 2 The level- and segment- based change-point models as particular cases of the HMM formalisms

We present a unifying HMM approach that can be adapted to many different approaches to the change-point problem. The known properties and current algorithms of the HMM allow for the efficient estimation of many quantities of interest. In turn, the results from the HMM may be used for model selection, or provide additional information about a given segmentation, such as the uncertainty of the change-points. The HMM framework also permits various extensions of the change-point model for further practical use.

A typical HMM is defined by the joint probability distribution

$$\mathbb{P}(X_{1:n}, S_{1:n}) = \mathbb{P}(S_1) \mathbb{P}(X_1|S_1) \prod_{i=2}^n \mathbb{P}(S_i|S_{i-1}) \mathbb{P}(X_i|S_i) \quad (2)$$

where  $n$  is the total number of observations,  $X_{1:n} = (X_1, \dots, X_n)$  is the vector of all the observation variables,  $S_{1:n} = (S_1, \dots, S_n)$  is the vector of all the hidden variables. Figure 1 provides a simple example of the dependencies among variables.

For *homogeneous* HMMs define  $\mathbb{P}(S_1 = s) = \mu(s)$ ,  $\mathbb{P}(S_i = s|S_{i-1} = r) = \alpha(r, s)$  for all  $i = 2, \dots, n$  and  $\mathbb{P}(X_i = x|S_i = s; \theta) = \beta_s(x)$  for all  $i = 1, \dots, n$ .

Given a *evidence*, or some prior knowledge of the states of some variables, the computation of the posterior probabilities of the hidden variables is an important aspect of Hidden Markov modeling. For all observations  $i = 1, \dots, n$ , we introduce the formal notion of *evidence* by considering  $\mathcal{X}_i$  and  $\mathcal{S}_i$ , which are subsets of the two sets of all possible outcomes of  $X_i$  and  $S_i$ , respectively. The

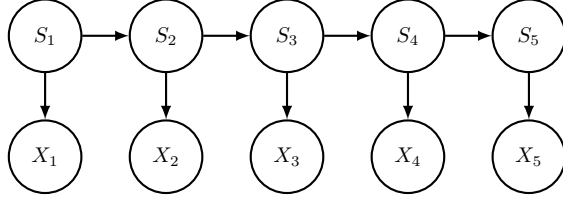


Figure 1: HMM topology with  $n = 5$ . For  $i = 1 \dots 5$ ,  $S_i$  are the hidden states, and  $X_i$  are the observed states.

evidence is

$$\mathcal{E} = \{X_i \in \mathcal{X}_i, S_i \in \mathcal{S}_i, \text{ for all } i = 1, \dots, n.\} \quad (3)$$

This evidence, which can comprise both observed information  $X_i$  and/or known prior information  $S_i$ , provides a constraint on the set of posterior distributions. The unconstrained case occurs when  $\mathcal{X}_i$  and  $\mathcal{S}_i$  both contain the set of all possible states for  $X_i$  and  $S_i$ , respectively, for all  $i = 1, \dots, n$ .

Depending on the definition of the hidden state variables  $S$ , their corresponding probability distributions and the available evidence, the HMM framework allows the definition of level- and segment-based models as follows.

## 2.1 Level-based model: the standard evidence

In the level-based model, the hidden state  $S_i$  pertains to the level (or underlying distribution) of observation  $X_i$ . This is an appropriate model when underlying properties can be shared between observations in different, non-adjacent segments. This HMM can be defined similarly to the classical HMM segmentation models by choosing the finite set of  $L \geq 1$  levels, with  $S \in \{1, 2, \dots, L\}^n$ . With this level-based approach  $K \geq L$ , and transitions are possible between any pair of states.

Because all observation variables  $X_i$  are observed and there are no constraints on the hidden states, the evidence is

$$\mathcal{E}^{\mathcal{L}} = \{X_1 = x_1, \dots, X_n = x_n\} = \{X_{1:n} = x_{1:n}\}. \quad (4)$$

This is the standard evidence usually found in most applications.

The transition matrix between hidden states can be assumed to be homogeneous, and is often parsimoniously parametrized. For example, consider:

$$\alpha(r, s) = \begin{cases} 1 - \eta_r & \text{if } s = r \\ \frac{\eta_r}{L-1} & \text{if } s \neq r \end{cases} \quad (5)$$

which uses only  $L$  free parameters.

## 2.2 Segment-based model

In the segment-based model, the objective is to find the best partitioning  $S \in \mathcal{M}_K$  of the data into  $K$  non-overlapping intervals, where the hidden state  $S_i$  pertains to the segment index of observation  $X_i$ .

One set of constraints [Luong et al., 2012] that allows the HMM to correspond *exactly* to the above segment-based model defined in Equation (1) is

$$\mathcal{E}^S = \{S_1 = 1, S_n = K, X_{1:n} = x_{1:n}\} \quad (6)$$

where the transition only permits increments of 0 or +1 of the segment index  $S_i$ , where  $K$  is the fixed total number of segments.

In order to obtain a uniform prior distribution of  $S_{1:n}$  over all possible segmentations with  $K$  segments  $\mathcal{M}_K$ , define the transition matrix over the state space  $\{1, \dots, K, K+1\}$  (where  $K+1$  is a “junk” state) as follows:

$$\alpha(r, s) = \begin{cases} 1 - \eta & \text{if } r \leq K \text{ and } s = r \\ \eta & \text{if } r \leq K \text{ and } s = r + 1 \\ 1 & \text{if } s = r = K + 1 \\ 0 & \text{otherwise} \end{cases}$$

where  $\eta \in ]0, 1[$  is a fixed number. Note that the particular value of  $\eta$  does affect  $\mathbb{P}(S)$  but has no effect whatsoever on  $\mathbb{P}(S|\mathcal{E}^S)$ . An arbitrary choice of  $\eta = 0.5$  is sufficient for practical computations.

### 3 Inference in level- and segment-based HMMs

#### 3.1 Variable elimination

The forward-backward algorithm [Rabiner, 1989, Durbin et al., 1998] efficiently solves many inference problems in HMMs; before going into its technical details we illustrate the motivation and the very simple ideas behind it.

Let us consider a HMM of length  $n$  in which each hidden variable has  $K$  possible states. Suppose we observe the standard evidence  $\mathcal{E} = \{X_{1:n} = x_{1:n}\}$ . This is the specific expression for evidence (4) found in the level-based model; similar expressions also apply to the evidence in the segment-based model (6) and for the most general form of evidence (3).

An important problem in inference is to estimate the probability of this evidence, which can be accomplished by summing the joint distribution  $\mathbb{P}(S_{1:n}, X_{1:n} = x_{1:n})$  over all the hidden variables:

$$\mathbb{P}(X_{1:n} = x_{1:n}) = \sum_{S_1, \dots, S_n} \mathbb{P}(S_1, \dots, S_n, X_{1:n} = x_{1:n}).$$

In a naive approach, the first step would be to evaluate the joint probability for each possible value of  $(S_1, \dots, S_n)$  and then perform the summations explicitly. However this is a highly inefficient method as its total cost is  $O(K^n)$ .

The exponential blowup is addressed by observing that the factors in the joint distribution (2) depend each on a small number of variables. A much more efficient algorithm is to evaluate expressions depending on these factors once and then cache the results, which avoids generating them multiple times. The basic tools which accomplish this are the factorization given in Eq. (2) and the distributive property.

For example, in the case  $n = 3$ :

$$\mathbb{P}(X_{1:n} = x_{1:n}) = \sum_{S_1} \mathbb{P}(S_1) \mathbb{P}(X_1 = x_1 | S_1) \underbrace{\sum_{S_2} \mathbb{P}(S_2 | S_1) \mathbb{P}(X_2 = x_2 | S_2) \underbrace{\sum_{S_3} \mathbb{P}(S_3 | S_2) \mathbb{P}(X_3 = x_3 | S_3)}_{B_2(S_2)}}_{B_1(S_1)}. \quad (7)$$

In practice, we choose an ordering of the hidden variables (here the *backward* order  $S_3 < S_2 < S_1$ ) and then rearrange all factors in order, so that all the factors depending on  $S_3$  are the furthest on the right. Then all remaining factors depending on  $S_2$  are placed to the left, and similarly for  $S_1$ . This makes it possible to eliminate one variable after another, according to the initial order. For each eliminated variable, we obtain a quantity which we cache and use in order to eliminate the ensuing variable. These quantities are also called *messages* and the outlined procedure provides a recursive method to compute them. In the case of the backward ordering, the messages are called *backward quantities*.

The computation of  $B_2(S_2)$  for all the values of  $S_2$  requires the sum of  $K$  terms (one for each value of  $S_3$ ); thus resulting in  $O(K^2)$  operations. Similarly, computing  $B_1$  requires  $O(K^2)$  operations. Because there are  $n$  hidden variables the resulting total cost is  $O(nK^2)$  which is a dramatic improvement over the naive  $O(K^n)$  complexity.

A central aspect of variable elimination is the initial ordering of these hidden variables. For instance, consider the elimination order  $S_2 < S_3 < S_1$ , then:

$$\mathbb{P}(X_{1:n} = x_{1:n}) = \sum_{S_1} \mathbb{P}(S_1) \mathbb{P}(X_1 = x_1 | S_1) \underbrace{\sum_{S_3} \mathbb{P}(X_3 = x_3 | S_3) \underbrace{\sum_{S_2} \mathbb{P}(S_3 | S_2) \mathbb{P}(S_2 | S_1) \mathbb{P}(X_2 = x_2 | S_2)}_{C(S_3, S_1)}}_{D(S_1)}.$$

The cost of eliminating  $S_2$  is  $O(K^3)$ , with  $K$  terms summed for each value of the pair  $(S_3, S_1)$ . As a result, the resulting complexity is  $O(nK^3)$ .

Another ordering which leads to a  $O(nK^2)$  cost is the *forward* ordering  $S_1 < S_2 < S_3$ :

$$\mathbb{P}(X_{1:n} = x_{1:n}) = \sum_{S_3} \mathbb{P}(X_3 = x_3 | S_3) \underbrace{\sum_{S_2} \mathbb{P}(S_3 | S_2) \mathbb{P}(X_2 = x_2 | S_2) \underbrace{\sum_{S_1} \mathbb{P}(S_2 | S_1) \mathbb{P}(X_1 = x_1 | S_1) \mathbb{P}(S_1)}_{E_2(S_2)}}_{E_3(S_3)}.$$

$E_2(S_2)$  and  $E_3(S_3)$  are closely related to the *forward* quantities defined in Section 3.2.1: having defined  $F_1(S_1) := \mathbb{P}(X_1 = x_1 | S_1) \mathbb{P}(S_1)$ :

$$E_2(S_2) = \sum_{S_1} \mathbb{P}(S_2 | S_1) F_1(S_1),$$

and

$$F_2(S_2) := \mathbb{P}(X_2 = x_2 | S_2) E_2(S_2) = \mathbb{P}(X_2 = x_2 | S_2) \sum_{S_1} \mathbb{P}(S_2 | S_1) F_1(S_1)$$

and

$$F_3(S_3) := \mathbb{P}(X_3 = x_3 | S_3) E_3(S_3) = \mathbb{P}(X_3 = x_3 | S_3) \sum_{S_2} \mathbb{P}(S_3 | S_2) F_2(S_2).$$

At this point there are two alternative formulae for solving our initial inference problem based on the recursive forward and backward quantities:

$$\mathbb{P}(X_{1:n} = x_{1:n}) = \sum_{S_3} F_3(S_3),$$

and also, from Eq. (7):

$$\mathbb{P}(X_{1:n} = x_{1:n}) = \sum_{S_1} F_1(S_1) B_1(S_1).$$

The next section explains the use of forward and backward quantities to compute the posterior probabilities  $\mathbb{P}(S_i | X_{1:n} = x_{1:n})$  and other useful distributions. For simplicity we will consider homogeneous HMMs.

## 3.2 Forward-backward algorithm

### 3.2.1 Level-based model: the standard forward-backward algorithm

In the level-based case, the evidence (4) is the standard evidence found in many HMMs applications. The inference problem is then solved by the classical recursive formulae used for most applications.

Define the forward and backward quantities:

$$F_i^{\mathcal{L}}(s) := \mathbb{P}(S_i = s, X_{1:i} = x_{1:i})$$

and

$$B_i^{\mathcal{L}}(s) := \mathbb{P}(X_{i+1:n} = x_{i+1:n} | S_i = s),$$

for all  $i \in \{1, \dots, n\}$  under the convention that  $B_n^{\mathcal{L}} \equiv 1$ . These probability distributions are exactly the standard forward-backward quantities found in most applications.

The recursive computation of the forward and backward quantities requires the following results:

**Proposition 1.** For all  $i = 1, \dots, n$  and for each value  $s$  of the hidden states:

$$\mathbb{P}(S_i = s, \mathcal{E}^{\mathcal{L}}) = F_i^{\mathcal{L}}(s) B_i^{\mathcal{L}}(s), \quad (8)$$

where  $\mathcal{E}^{\mathcal{L}}$  is defined in (4). Moreover for all  $i \in \{2, \dots, n\}$ , for each pair  $(r, s)$  and for each observation  $x$ :

$$\mathbb{P}(S_{i-1} = r, S_i = s, \mathcal{E}^{\mathcal{L}}) = F_{i-1}^{\mathcal{L}}(r) \alpha(r, s) \beta_s(x) B_i^{\mathcal{L}}(s) \quad (9)$$

*Proof.* We start by proving Eq. (8), by applying the chain rule and the conditional independence of the events  $\{X_{i+1:n} = x_{i+1:n}\}$  and  $\{X_{1:i} = x_{1:i}\}$  given  $\{S_i = s\}$ :

$$\begin{aligned} \mathbb{P}(S_i = s, \mathcal{E}^{\mathcal{L}}) &= \mathbb{P}(S_i = s, X_{1:i} = x_{1:i}, X_{i+1:n} = x_{i+1:n}) = \\ &= \mathbb{P}(S_i = s, X_{1:i} = x_{1:i}) \mathbb{P}(X_{i+1:n} = x_{i+1:n} | S_i = s, X_{1:i} = x_{1:i}) = \\ &= \mathbb{P}(S_i = s, X_{1:i} = x_{1:i}) \mathbb{P}(X_{i+1:n} = x_{i+1:n} | S_i = s). \end{aligned}$$

Similarly for Equation (9):

$$\begin{aligned}\mathbb{P}(S_{i-1} = r, S_i = s, \mathcal{E}^{\mathcal{L}}) &= \mathbb{P}(S_{i-1} = r, X_{1:i-1} = x_{1:i-1}, S_i = s, X_i = x_i, X_{i+1:n} = x_{i+1:n}) = \\ &= \mathbb{P}(S_{i-1} = r, X_{1:i-1} = x_{1:i-1}) \mathbb{P}(S_i = s | S_{i-1} = r) \times \\ &= \mathbb{P}(X_i = x_i | S_i = s) \mathbb{P}(X_{i+1:n} = x_{i+1:n} | S_i = s).\end{aligned}$$

□

This proposition establishes all the classical results for inference in HMMs.

**Corollary 2** (Forward and backward recursions). The forward quantities can be computed iteratively starting from  $F_1^{\mathcal{L}}(s) = \mu(s)\beta_s(x_1)$  for all  $i = 2, \dots, n$  with

$$F_i^{\mathcal{L}}(s) = \sum_r F_{i-1}^{\mathcal{L}}(r) \alpha(r, s) \beta_s(x_i). \quad (10)$$

The backward quantities can be computed recursively from  $B_n^{\mathcal{L}} \equiv 1$  for all  $i = n-1, \dots, 1$  with

$$B_{i-1}^{\mathcal{L}}(r) = \sum_s \alpha(r, s) \beta_s(x_i) B_i^{\mathcal{L}}(s). \quad (11)$$

*Proof.* In order to obtain the forward Equation (10), apply Equations (8) and (9) respectively to the right and left hand side of

$$\mathbb{P}(S_i = s, \mathcal{E}^{\mathcal{L}}) = \sum_r \mathbb{P}(S_{i-1} = r, S_i = s, \mathcal{E}^{\mathcal{L}}).$$

A similar argument holds for the backward Equation (11). □

The following useful result is a straightforward consequence of Proposition 1:

**Corollary 3** (Posterior probabilities). For each  $i = 1, \dots, n$ , the posterior state probabilities are

$$\mathbb{P}(S_i = s | \mathcal{E}^{\mathcal{L}}) = \frac{F_i^{\mathcal{L}}(s) B_i^{\mathcal{L}}(s)}{\mathbb{P}(\mathcal{E}^{\mathcal{L}})} \quad \text{and} \quad \mathbb{P}(S_{i-1} = r, S_i = s | \mathcal{E}^{\mathcal{L}}) = \frac{F_{i-1}^{\mathcal{L}}(r) \alpha(r, s) \beta_s(x_i) B_i^{\mathcal{L}}(s)}{\mathbb{P}(\mathcal{E}^{\mathcal{L}})}$$

where the probability of the evidence is

$$\mathbb{P}(\mathcal{E}^{\mathcal{L}}) = \sum_s F_i^{\mathcal{L}}(s) B_i^{\mathcal{L}}(s),$$

for an arbitrary fixed  $i \in \{1, \dots, n\}$ ; in particular  $\mathbb{P}(\mathcal{E}^{\mathcal{L}}) = \sum_s F_n^{\mathcal{L}}(s)$ .

The following corollary is easily proven and makes it possible to sample the joint distribution of the hidden variables recursively:

**Corollary 4** (Forward and backward sampling). The joint distribution of  $S_{1:n}$  conditionally to  $\mathcal{E}^{\mathcal{L}}$  is a Markov chain whose transitions are given by

$$\mathbb{P}(S_i = s | S_{i-1} = r, \mathcal{E}^{\mathcal{L}}) = \frac{\alpha(r, s) \beta_s(x_i) B_i^{\mathcal{L}}(s)}{B_{i-1}^{\mathcal{L}}(r)}$$

in the forward direction, and by

$$\mathbb{P}(S_{i-1} = r | S_i = s, \mathcal{E}^{\mathcal{L}}) = \frac{F_{i-1}^{\mathcal{L}}(r) \alpha(r, s) \beta_s(x_i)}{F_i^{\mathcal{L}}(s)}$$

in the backward direction.

### 3.2.2 Forward-backward algorithm for segment-based model

For the sake of completeness, we present the results of the inference problem for the HMMs constrained by the evidence (6). The formulae presented here are easily obtained by modifying the standard formulae proved in section 3.2.1 and are particular cases of the forward-backward algorithm for HMMs conditioned on the general evidence (3). In particular, the formulae from the previous section still hold for  $i \notin \{1, n\}$ . In these two special cases consider the constraints  $S_1 = 1$  and  $S_n = K$ , which are imposed by adding the multiplicative constants  $\mathbf{1}_{\{S_1=1\}}$  or  $\mathbf{1}_{\{S_n=K\}}$  to the formulae<sup>1</sup>.

Define the forward quantities as

$$F_i^{\mathcal{S}}(s) := \mathbb{P}(S_1 = 1, S_i = s, X_{1:i} = x_{1:i})$$

for  $i \leq n - 1$  and

$$F_n^{\mathcal{S}}(s) := \mathbb{P}(S_1 = 1, S_n = s, S_n = K, X_{1:n} = x_{1:n}) = \mathbf{1}_{\{s=K\}} \mathbb{P}(S_1 = 1, S_n = s, X_{1:n} = x_{1:n}).$$

Define the backward quantities as

$$B_i^{\mathcal{S}}(s) := \mathbb{P}(X_{i+1:n} = x_{i+1:n}, S_n = K | S_i = s),$$

for all  $i \in \{1, \dots, n\}$ , with the convention that  $B_n^{\mathcal{S}} \equiv 1$ . The corresponding equations for (8) and (9) become  $\mathbb{P}(S_i = s, \mathcal{E}^{\mathcal{S}}) = F_i^{\mathcal{S}}(s) B_i^{\mathcal{S}}(s)$  and

$$\mathbb{P}(S_{i-1} = r, S_i = s, \mathcal{E}^{\mathcal{S}}) = F_{i-1}^{\mathcal{S}}(r) \alpha(r, s) \beta_s(x) B_i^{\mathcal{S}}(s)$$

for all  $i = 1, \dots, n - 1$  and

$$\mathbb{P}(S_{n-1} = r, S_n = s, \mathcal{E}^{\mathcal{S}}) = \mathbf{1}_{\{s=K\}} F_{n-1}^{\mathcal{S}}(r) \alpha(r, s) \beta_s(x) B_n^{\mathcal{S}}(s).$$

The forward quantities can be computed iteratively starting from  $F_1^{\mathcal{S}}(s) = \mathbf{1}_{\{s=1\}} \mu(s) \beta_s(x_1)$  for all  $i = 2, \dots, n - 1$  with

$$F_i^{\mathcal{S}}(s) = \sum_r F_{i-1}^{\mathcal{S}}(r) \alpha(r, s) \beta_s(x_i),$$

and  $F_n^{\mathcal{S}}(s) = \mathbf{1}_{\{s=K\}} \sum_r F_{n-1}^{\mathcal{S}}(r) \alpha(r, s) \beta_s(x_i)$ . The recursions for computing the backward quantities are exactly the same as in the level-based setting:

$$B_{i-1}^{\mathcal{S}}(r) = \sum_s \alpha(r, s) \beta_s(x_i) B_i^{\mathcal{S}}(s).$$

The preceding constraints result in less computations due to the sparse transition matrix between hidden states, as  $\alpha(r, s) = 0$  if  $s - r \notin (0, 1)$  leading to a  $O(Kn)$  complexity. In the constrained model we note that the probability of the evidence can be computed, for instance, with  $\mathbb{P}(\mathcal{E}^{\mathcal{S}}) = F_1^{\mathcal{S}}(1) B_1^{\mathcal{S}}(1)$ .

---

<sup>1</sup>Given an event  $\mathcal{A}$ ,  $\mathbf{1}_{\mathcal{A}} = 1$  iff  $\mathcal{A}$  is true.



### 3.3 Extensions

Extending the previously described recursive formulae and results to more general evidence forms is a straightforward process. For instance, suppose that we observe  $X_i = x_i$  for all  $i$  except for  $i = 5$ , which we will consider as missing. In this situation, the evidence is  $\mathcal{E} = \{X_{1:4} = x_{1:4}, X_{6:n} = x_{6:n}\}$ . Define the forward and backward quantities exactly as in the level-based case with the only difference that for  $i \geq 5$ ,  $F_i(s) = \mathbb{P}(S_i = s, X_{1:4} = x_{1:4}, X_{6:i} = x_{6:i})$  and for  $i \leq 4$   $B_i(s) = \mathbb{P}(X_{i+1:4} = x_{i+1:4}, X_{6:n} = x_{6:n} | S_i = s)$ . In practice, all the results from the level-based case still hold if we substitute  $\beta_s(x_5)$  by 1 in the formulae expressing  $\mathbb{P}(S_4 = r, S_5 = s, \mathcal{E})$ ,  $F_5(s)$ ,  $B_4(r)$ ,  $\mathbb{P}(S_5 = s | S_4 = r, \mathcal{E})$  and  $\mathbb{P}(S_4 = r | S_5 = s, \mathcal{E})$ .

Another situation is all variables  $(X_1, \dots, X_n)$  being observed with additional partial prior knowledge about the hidden variables. Suppose for instance  $X_{1:n} = x_{1:n}$  and, say,  $S_7 \neq 2$ . The evidence is  $\mathcal{E} = \{X_{1:n} = x_{1:n}, S_7 \neq 2\}$  and we define the forward and backward quantities as before with the following exceptions:  $F_i(s) = \mathbb{P}(X_{1:i} = x_{1:i}, S_7 \neq 2, S_i = s)$  for  $i \geq 7$  and  $B_i(s) = \mathbb{P}(X_{i+1} = x_{i+1}, S_7 \neq 2 | S_i = s)$  for  $i < 7$ . All the results continue to hold by substituting  $\alpha(r, s)$  by  $\mathbf{1}_{s \neq 2} \cdot \alpha(r, s)$  in the formulae expressing  $\mathbb{P}(S_6 = r, S_7 = s, \mathcal{E})$ ,  $F_7(s)$ ,  $B_6(r)$ ,  $\mathbb{P}(S_7 = s | S_6 = r, \mathcal{E})$  and  $\mathbb{P}(S_6 = r | S_7 = s, \mathcal{E})$ .

The forward-backward algorithm is also known as the *product-sum algorithm* and is a particular case of the *message propagation* algorithm for Bayesian networks [Koller and Friedman, 2009]. As previously seen, its basic mechanism is the simple distributive property of multiplication over addition.

Another common inferential problem in HMMs is in finding a set of variables with the largest joint state probability and their corresponding marginal probabilities. In general, the set which maximizes the joint state probabilities may not exactly match the set that maximizes each marginal probability separately. The *max-sum* algorithm (also known as the *Viterbi* algorithm) addresses this issue by replacing summations by maximizations in all the above formulae [Viterbi, 1967, Rabiner, 1989]. The previous results continue to hold because the multiplication is distributive over the max operator:  $\max(ab, ac) = a \max(b, c)$ .

### 3.4 Floating-point issues

Underflow is a common issue when using forward and/or backward recursions in floating-point arithmetic. Indeed, the magnitude of the forward quantity  $F_i$  decreases geometrically with  $i$  and can be smaller than the smallest machine float (ex:  $2.23 \times 10^{-308}$  for C++ double-precision on i686 architecture) which is an architecture-dependent threshold. As suggested in [Rabiner, 1989, pages 272-273], one solution consists in keeping track of a rescaling parameter (typically stored in log-scale) for each  $i$ . To improve this inefficient approach, in terms of both time and memory, we suggest to use log-scale computation through  $\log F_i$  and  $\log B_i$  both for level- and segment-based models.

For both expressions of  $F_i$  and  $B_i$  we recommend the initial precomputation of  $\log \alpha(r, s)$  and  $\log \beta_s(x_i)$  for all  $r, s, i$ . The forward and backward recursions become:

$$\log F_i(s) = \text{logsum}[\log F_{i-1}(\cdot) + \log \alpha(\cdot, s) + \log \beta_s(x_i)]$$

$$\log B_{i-1}(r) = \text{logsum}[\log \alpha(r, \cdot) + \log \beta_\cdot(x_i) + \log B_i(\cdot)]$$

where  $\text{logsum}$  is a function of any real vector  $z = (z_1, z_2, \dots, z_k)$ . For example if  $z_1 \leq z_2 \leq \dots \leq z_k$

and  $\text{logsum}(z_1, z_2, \dots, z_k) = \log \sum_j \exp(z_j)$ :

$$\text{logsum}(z_1, z_2, \dots, z_k) = z_1 + \text{log1p}[\exp(z_2 - z_1) + \dots + \exp(z_k - z_1)]$$

where  $\text{log1p}(u) = \log(1 + u)$ , which is useful in floating-point arithmetic when  $u$  is small.

### 3.5 E-M algorithm

Let  $\mathcal{E}$  be either  $\mathcal{E}^{\mathcal{L}}$  or  $\mathcal{E}^{\mathcal{S}}$  or a more general evidence. The Expectation-Maximization (EM) algorithm consists of repeating iteratively the two following steps:

- *Expectation (E)*: compute  $Q(\theta'|\theta) = \sum_S \mathbb{P}_\theta(S|\mathcal{E}) \log \mathbb{P}_{\theta'}(S, \mathcal{E})$ ;
- *Maximization (M)*: update  $\theta$  with  $\tilde{\theta} = \arg \max_{\theta'} Q(\theta'|\theta)$ .

The resulting update formulae depend on the model considered (level- or segment-based) and on the nature of the emission distribution. In this section, we consider only the two following emission models:

- *normal homoscedastic*: for all  $x \in \mathbb{R}$  and hidden state  $s$ ,  $\beta_s(x) = \varphi((x - \mu_s)/\sigma)$  where  $\mu_s \in \mathbb{R}$ ,  $\sigma > 0$ , with  $\varphi(z) = \exp(z^2/2)/\sqrt{2\pi}$  for all  $z \in \mathbb{R}$ ;
- *Poisson*: for all  $k \in \mathbb{N}$  and hidden state  $s$ ,  $\beta_s(k) = \exp(-\lambda_s)\lambda_s^k/k!$  where  $\lambda_s > 0$ .

During the E-step, forward-backward recursions obtain  $\mathbb{P}_\theta(S|\mathcal{E})$  as a heterogeneous Markov chain. In practice, it is only necessary to compute  $\mathbb{P}_\theta(S_i|\mathcal{E}) = F_i(S_i)B_i(S_i)/\mathbb{P}_\theta(\mathcal{E})$  and  $\mathbb{P}_\theta(S_{i-1}, S_i|\mathcal{E}) = F_{i-1}(S_{i-1})\alpha(S_{i-1}, S_i)\beta_{S_i}(x_i)B_i(X_i)/\mathbb{P}_\theta(\mathcal{E})$ .

For both level- and segment- based model, the M-step is the same for the emission part of  $Q(\theta'|\theta)$ :

$$\tilde{\mu}_s = \tilde{\lambda}_s = \frac{\sum_{i=1}^n X_i \mathbb{P}_\theta(S_i = s|\mathcal{E})}{\sum_{i=1}^n \mathbb{P}_\theta(S_i = s|\mathcal{E})} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{\sum_{i=1}^n \sum_s (X_i - \tilde{\mu}_s)^2 \mathbb{P}_\theta(S_i = s|\mathcal{E})}{n}.$$

For the transition part, there is no parameter to update in the segment-based case, and for the level-based case (with the transition matrix defined in Eq.5):

$$\tilde{\eta}_r = \frac{\sum_{r \neq s} \sum_{i=2}^n \mathbb{P}_\theta(S_{i-1} = r, S_i = s|\mathcal{E})}{\sum_{i=2}^n \mathbb{P}_\theta(S_{i-1} = r|\mathcal{E})}.$$

### 3.6 Change-point estimation

#### 3.6.1 Level-based model

Change-points are identified based on observations where transitions are most likely to occur. In a HMM framework, we assess the probability that  $i$  is any change-point between two different hidden states, regardless of the ordering within the set of change-points.

The posterior probability of any change-point occurring at observation  $i$ , or such that  $S_i \neq S_{i+1}$ , is:

$$\begin{aligned}\mathbb{P}(CP = i|\mathcal{E}^{\mathcal{L}}) &= \sum_{r \neq s} \mathbb{P}(S_i = r, S_{i+1} = s|\mathcal{E}^{\mathcal{L}}) \\ &= \sum_{r \neq s} \mathbb{P}(S_{i+1} = s|S_i = r, \mathcal{E}^{\mathcal{L}}) \mathbb{P}(S_i = r|\mathcal{E}^{\mathcal{L}}) \\ &= \sum_{r \neq s} \frac{F_i^{\mathcal{L}}(r) \alpha(r, s) \beta_s(x_{i+1}) B_{i+1}^{\mathcal{L}}(s)}{F_1^{\mathcal{L}}(1) B_1^{\mathcal{L}}(1)}\end{aligned}$$

for  $i = 1, \dots, n-1$  and  $r, s = 1, \dots, L$ , where the last equality follows from Section 3.2.1.

Unlike the segment-based approach, the level-based approach does not allow to obtain the distribution of the specific  $r^{th}$  change-point, but only the marginal probability of a change-point at a given position. It is however possible to derive the distribution of the  $r^{th}$  change-point in the level-based approach as the waiting time of a regular expression in a heterogeneous Markov chain [Aston and Martin, 2007].

### 3.6.2 Segment-based model

Define the location of the  $r^{th}$  change-point as  $CP_r$ . Under this definition, the posterior probability of the  $r^{th}$  change at observation  $i$  is  $\mathbb{P}(CP_r = i) = \mathbb{P}(S_i = r, S_{i+1} = s|\mathcal{E}^{\mathcal{S}})$ , where  $s = r + 1$ . In other words the  $r^{th}$  change-point is the last observation of segment  $r$  before segment  $r + 1$ , each of whom consists of contiguous, homogeneous observations.

Using the formulae provided in Section 3.2.2, the posterior probability of the  $r^{th}$  change-point occurring after observation  $i$ , where  $s = r + 1$  is:

$$\begin{aligned}\mathbb{P}(CP_r = i|\mathcal{E}^{\mathcal{S}}) &= \mathbb{P}(S_i = r, S_{i+1} = s|\mathcal{E}^{\mathcal{S}}) \\ &= \mathbb{P}(S_{i+1} = s|S_i = r, \mathcal{E}^{\mathcal{S}}) \mathbb{P}(S_i = r|\mathcal{E}^{\mathcal{S}}) \\ &= \frac{F_i^{\mathcal{S}}(r) \alpha(r, r+1) \beta_{r+1}(x_{i+1}) B_{i+1}^{\mathcal{S}}(r+1)}{F_1^{\mathcal{S}}(1) B_1^{\mathcal{S}}(1)}.\end{aligned}$$

for  $i = 1, \dots, n-1$  and  $r = 1, \dots, K-1$ .

## 4 Examples

### 4.1 British coal mining disaster data set

We illustrate the methods on the classical British coal mining disaster data set [Carlin et al., 1992], which displays the number of accidents per year in Great Britain between 1851 and 1962,  $n = 112$ . The observed count data in the HMM use a Poisson emission distribution. For a change-point model with 3 segments, a greedy least squares minimization algorithm [Hartigan and Wong, 1979] identified change-points at  $i = 36$  and  $97$ , corresponding to the years 1886 and 1947, respectively.

The level-based approach models the change-points as transitions between states, which correspond to mean parameters of the Poisson distribution, or levels. Using these initial change-points, we obtain maximum likelihood estimates of means  $\hat{\lambda} = (3.25, 1.15, 0.27)$  and the transition probabilities to be  $\hat{\eta} = (1/36, 1/61)$ . Through the forward-backward algorithm described in section 3.2.1

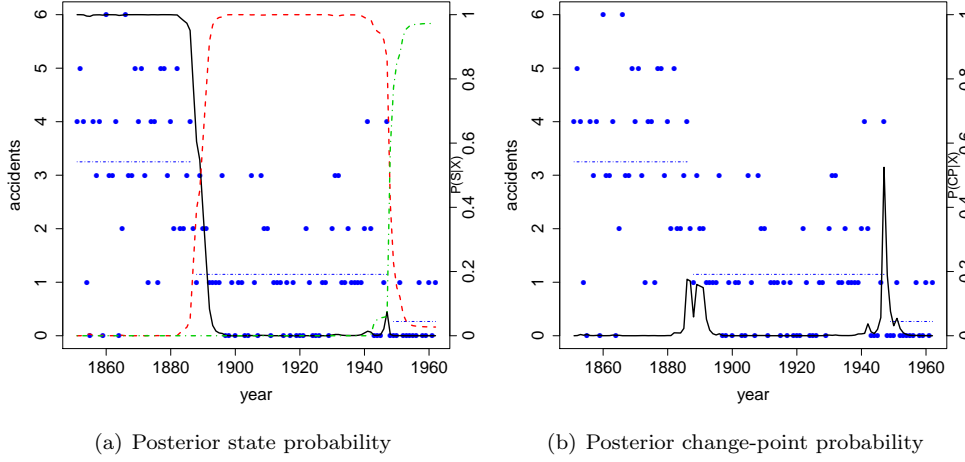


Figure 2: *Level-based model*. Plots of estimated posterior probabilities in British coal mining disaster data set [Carlin et al., 1992]. Dots are annual accidents with horizontal lines being the maximum likelihood estimates of the  $L = 3$  level means. (a) Posterior marginal probability of  $i$  being in state  $r$ ,  $\mathbb{P}(S_i = r|\mathcal{E}^L)$ , with solid line being state 1, dashed line state 2 and dashed-dotted line state 3. (b) Posterior probability of  $i$  being any change-point  $\mathbb{P}(CP = i|\mathcal{E}^L) = \mathbb{P}(S_i = r, S_{i+1} = s|\mathcal{E}^L)$ , where  $r \neq s$ .

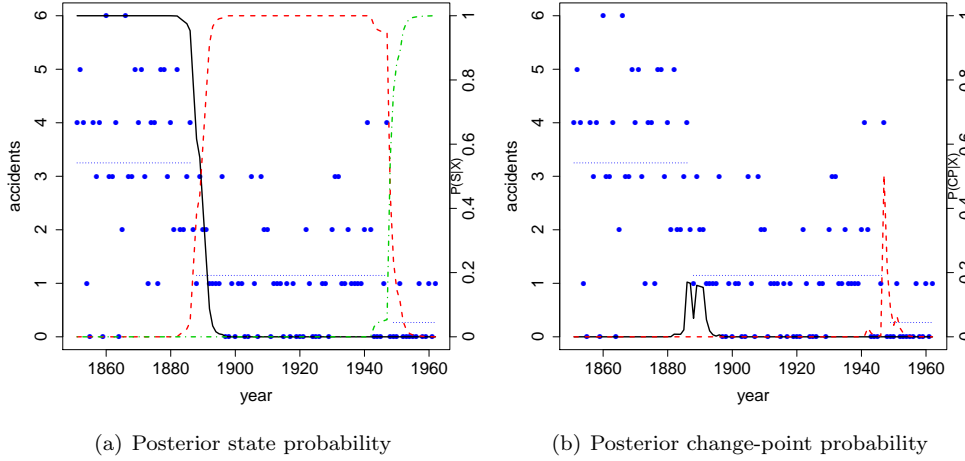


Figure 3: *Segment-based model*. Plots of estimated posterior probabilities in British coal mining disaster data set [Carlin et al., 1992]. Dots are annual accidents with horizontal lines being the maximum likelihood estimates of the  $K = 3$  segment means. (a) Posterior marginal probability of  $i$  being in state  $r$ ,  $\mathbb{P}(S_i = r|\mathcal{E}^S)$ , with solid line being segment 1, dashed line segment 2 and dashed-dotted line segment 3. (b) Posterior probability of  $i$  being  $r^{th}$  change-point  $\mathbb{P}(CP_r = i|\mathcal{E}^S) = \mathbb{P}(S_i = r, S_{i+1} = r + 1|\mathcal{E}^S)$ , with solid line being change-point 1, dashed line change-point 2.

the posterior estimates of the state of observation  $i$  given the data are  $\mathbb{P}(S_i = r|\mathcal{E}^{\mathcal{L}})$  (Figure 2(a)) for each of  $L = 3$  levels, and for any change occurring at  $i$   $\mathbb{P}(CP = i|\mathcal{E}^{\mathcal{L}})$  (Figure 2(b)).

Figures 3(a) and 3(b) display the posterior probabilities of the marginal distribution  $\mathbb{P}(S_i = r|\mathcal{E}^{\mathcal{S}})$  for each of  $K = 3$  segments and change-point probability  $\mathbb{P}(CP_r = i|\mathcal{E}^{\mathcal{S}}) = \mathbb{P}(S_i = r, S_{i+1} = r + 1|\mathcal{E}^{\mathcal{S}})$  for  $K - 1 = 2$  change-points, respectively. The segment-based approach models the change-points as the beginning and end of 3 intervals of contiguous observations with homogeneous distributions. Considering the hidden states of both level and segment -based HMMs are the same, the two different approaches produce almost identical respective marginal state and change-point curves.

From a historical perspective it is practical to look for events that may have triggered these detected changes in accident rates. In previous change-point literature, Raftery and Akman [1986] noted that the first change at year 1886 occurred during a decline in labor productivity at the end of the 1800s and the emergence of the Miners' Federation. Though, the uncertainty in the first change-point location suggests that the reduction in accidents was not due to a sudden event but to a continual decline over this time period. Meanwhile, the second change-point at 1947 coincides with a more clearly-defined time point that is two years after the end of the second World War.

## 4.2 Breast cancer data set

We apply the methods to a widely referenced data set from a breast cancer cell line BT474 [Snijders et al., 2001]. The data consist of log-reference ratios (LRRs) signifying the ratio of genomic copies of test samples compared to normal. The goal is to segment the data into segments with similar copy numbers, with change-points pointing to a copy number aberration that may signify genetic mutations of interest [Pinkel et al., 1998]. We use the same data previously analyzed [Rigaill et al., 2011], consisting of  $n = 120$  observations from chromosome 10. The observations are sorted according to their relative position along chromosome 10. The observed LRR data in the HMM uses a normal emission distribution.

For a change-point model with 4 segments, the least squares algorithm identified the most likely change-points at  $i = 68, 80$  and  $96$ . Under this specific level-based model, the observations in segments 1 and 3 may share the same distribution. Using these initial change-points, the maximum likelihood estimates are  $\hat{\mu} = (0.271, -0.039, -0.636)$  for the 3 levels and the transition probabilities to be  $\hat{\eta} = (2/84, 1/16)$ . We obtain the posterior estimates of the state of observation  $i$  given the data  $\mathbb{P}(S_i = r|\mathcal{E}^{\mathcal{L}})$  (Figure 4(a)) for  $L = 3$  levels and any change occurring at  $i$   $\mathbb{P}(CP = i|\mathcal{E}^{\mathcal{L}})$  (Figure 4(b)).

Figures 5(a) and 5(b) display the posterior probabilities of the marginal distribution  $\mathbb{P}(S_i = r|\mathcal{E}^{\mathcal{S}})$  for  $K = 4$  segments and change-point probabilities  $\mathbb{P}(CP_r = i|\mathcal{E}^{\mathcal{S}}) = \mathbb{P}(S_i = r, S_{i+1} = r + 1|\mathcal{E}^{\mathcal{S}})$  for  $K - 1 = 3$  change-points, respectively. The maximum likelihood estimates of means are  $\hat{\mu} = (0.289, -0.039, 0.224, -0.636)$  for the 4 segments. With different segments 1 and 3 defined as homogeneous, the two approaches produce slightly different probability distributions of change-point locations, though the locations of peaks remain similar.

Figure 6 compares the change-point probability curves for both  $K = 3$  and  $K = 4$  segment-based change-point models. The  $K = 3$  segment-based model does not include the second change-point at  $i = 80$ . The shape of the posterior probability curve of the first change-point slightly changes between the two models, due to the uncertainty in the precise location of this point. On the other hand, the peak of the last change-point is close to one, and due to this high precision, the corresponding change-point curve from both the  $K = 3$  and  $K = 4$  change-point models virtually

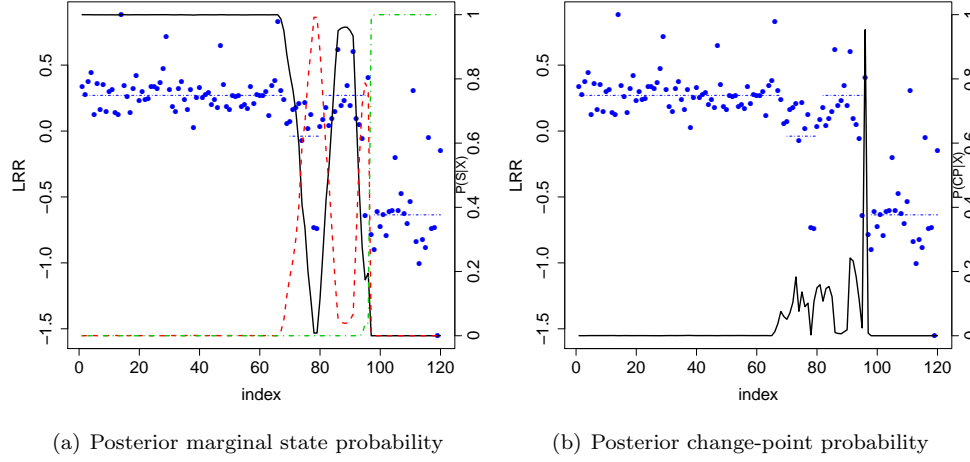


Figure 4: *Level-based model*. Plots of estimated posterior probabilities in breast-cancer data set [Snijders et al., 2001]. Dots are LRR with horizontal lines being the maximum likelihood estimates of the  $L = 3$  level means. (a) Posterior marginal probability of  $i$  being in state  $r$ ,  $\mathbb{P}(S_i = r|\mathcal{E}^{\mathcal{L}})$ , with solid line being state 1, dashed line state 2 and dashed-dotted line state 3. (b) Posterior probability of  $i$  being any change-point  $\mathbb{P}(CP = i|\mathcal{E}^{\mathcal{L}}) = \mathbb{P}(S_i = r, S_{i+1} = s|\mathcal{E}^{\mathcal{L}})$ , where  $r \neq s$ .

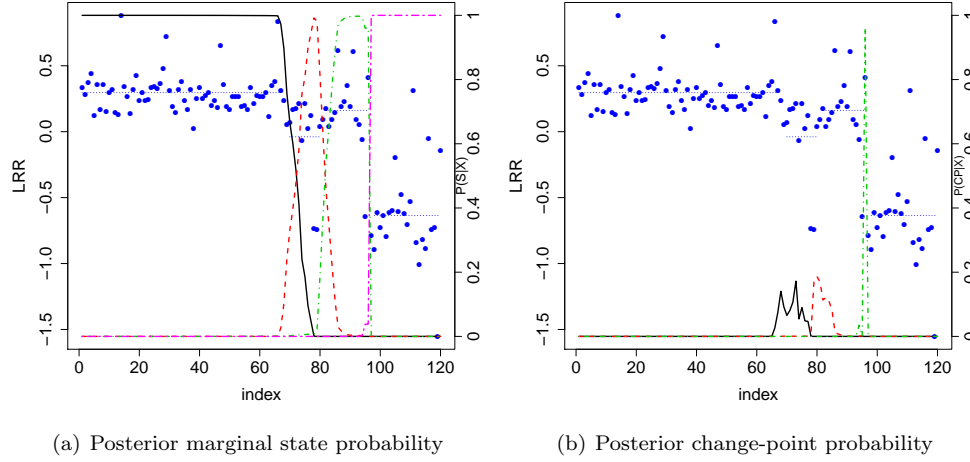


Figure 5: *Segment-based model*. Plots of estimated posterior probabilities in breast-cancer data set [Snijders et al., 2001]. Dots are LRR with horizontal lines being the maximum likelihood estimates of the  $K = 4$  segment means. (a) Posterior marginal probability of  $i$  being in state  $r$   $\mathbb{P}(S_i = r|\mathcal{E}^{\mathcal{S}})$ , with solid line being segment 1, dashed line segment 2, dashed-dotted line segment 3 and long dashed line segment 4. (b) Posterior probability of  $i$  being  $r^{th}$  change-point  $\mathbb{P}(CP_r = i|\mathcal{E}^{\mathcal{S}}) = \mathbb{P}(S_i = r, S_{i+1} = r + 1|\mathcal{E}^{\mathcal{S}})$ , with solid line being change-point 1, dashed line change-point 2, and dotted-dashed line change-point 3.

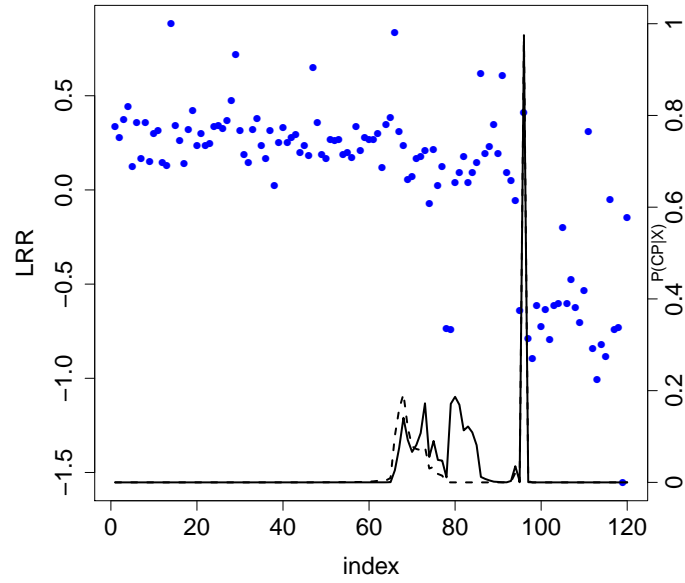


Figure 6: Plots of estimated posterior change-point probabilities of  $K = 3$  and  $K = 4$  segment change-point models in breast-cancer data set [Snijders et al., 2001]. Dots are LRR, lines are posterior probabilities of a change-point, where  $\mathbb{P}(CP_r = i | \mathcal{E}^S) = \mathbb{P}(S_i = r, S_{i+1} = r + 1 | \mathcal{E}^S)$ . Change-point probability curve is dotted for  $K = 3$  segment model and solid for  $K = 4$  segment model.

overlap.

## 5 Current implementations of change-point models for genomics data

A common point of interest in bioinformatics is to find genetic mutations pointing to phenotypes susceptible in cancer and other diseases. The detection of change points in Copy Number Variation (CNV) is a critical step in the characterization of DNA, including tumoral DNA in cancer. A CNV may locate a genetic mutation such as a duplication or deletion in a cancerous cell that is a target for treatment.

Many level-based approaches use the hidden state space in a classical discrete HMM to characterize mutations [Fridlyand et al., 2004] to map the number of states and the most likely state at each position. Various extensions to this HMM approach include various procedures such as merging change-points and specifying prior transition matrices [Willenbrock and Fridlyand, 2005, Marioni et al., 2006] to improve the results. Other extensions include reversible-jump Markov chain Monte Carlo (MCMC) to fit the HMM [Rueda and Díaz-Uriarte, 2007], a continuous-index HMM [Stjernqvist et al., 2007] that takes into account the discrete nature of observed genomics data, and methods that take into account genomic distance and overlap between clones [Andersson et al., 2008]. This HMM approach has also been extended for simultaneous change-points across multiple samples [Shah et al., 2009] and for simultaneously identifying multiple outcomes [Liu et al., 2010]. Other HMM-based implementations [Colella et al., 2007, Wang et al., 2007] deal with higher-resolution data for current SNP array technologies.

A wide amount of non-HMM approaches are also available for bioinformatics data. The aim of these approaches is typically in finding contiguous segments consisting of observations with the same distribution. One such implementation is a non-parametric extension of binary segmentation for multiple change-points through permutation [Olshen et al., 2004] along with a faster extension [Venkatraman and Olshen, 2007] which uses stopping rules. Various smoothing techniques [Hupé et al., 2004, Eilers and De Menezes, 2005, Hsu et al., 2005] have also been applied to change-point modelling. Summaries and comparisons of the various approaches for finding change-points in genomics data are available [Willenbrock and Fridlyand, 2005, Lai et al., 2005].

## 6 Conclusion

This chapter describes a simple algorithm using hidden Markov models to estimate posterior distributions of interest in change-point analysis, using two different modelling approaches. It also addresses computational issues through several simple constraints on the HMM to allow for estimates in a feasible amount of time.

HMMs are a special case of Bayesian Networks (BNs), which are probabilistic graphical models that represent random variables and their dependencies via a Directed Acyclic Graph (DAG), for example in Figure 1. The conditional dependencies coded by the edges of the DAG determine a factorization of the joint probability distribution. Given any type of evidence, summing variables out in the joint distribution obtains the conditional distributions of the variables. For BNs, this is done in a precise fashion by the exact Belief Propagation (BP) algorithm, which efficiently applies the distributive law to provide a recursive decomposition of the initial sums of products. The



intermediate sums of products (the so-called messages or beliefs) propagate through a secondary tree-shaped structure called a *junction tree* or *graph tree*. This general framework allows for recursive algorithms to obtain various quantities of interest including posterior distribution, likelihood or entropy-related terms. Moreover, the recursive algorithms for BNs permit model selection through the use of typical criteria such as the Bayesian Information Criterion [BIC, see Schwarz, 1978] or the Integrated Completed Likelihood [ICL, see Biernacki et al., 2000].

The exact BP general algorithm makes it possible to easily derive forward and backward recursions for HMMs for any kind of evidence. Moreover the BN unifying framework permits simple extensions of the aforementioned change-point HMMs to more complex and realistic models accounting for intricate dependencies between the variables or data. As an example, a straightforward generalization of the aforementioned segment- and level-based models is to account simultaneously for both hidden segments and levels. More specifically, let  $S_i$  and  $L_i$  be the segment and level, respectively, of observation  $X_i$ . In this specification, the current level  $L_i$  not only depends on  $L_{i-1}$ , but also on both the hidden states  $S_i$  and  $S_{i-1}$ .

Another useful extension is a model with joint segmentation of multiple samples. In a two-sample situation, there are two sets of distinct observations  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_n$ , one for each sample. Let  $U_1, \dots, U_n$  be the hidden variables corresponding to the segments of the first sample and  $V_1, \dots, V_n$  the variables corresponding to the segments of the second sample. The model assumes that  $U_1, \dots, U_n$  and  $V_1, \dots, V_n$  both depend on a common segmentation  $S_1, \dots, S_n$ . An advantage of this model is that the joint segmentation of both samples is not independent and can identify common features of both samples.

## References

- R. Andersson, C.E.G. Bruder, A. Piotrowski, U. Menzel, H. Nord, J. Sandgren, T.R. Hvidsten, T.D. de Ståhl, J.P. Dumanski, and J. Komorowski. A segmental maximum a posteriori approach to genome-wide copy number profiling. *Bioinformatics*, 24(6):751–758, 2008.
- J.A.D. Aston and D.E.K. Martin. Distributions associated with general runs and patterns in hidden Markov models. *The Annals of Applied Statistics*, pages 585–611, 2007.
- C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, 2000.
- J.A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4:126, 1998.
- O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Verlag, 2005.
- B.P. Carlin, A.E. Gelfand, and A.F.M. Smith. Hierarchical Bayesian analysis of changepoint problems. *Applied statistics*, pages 389–405, 1992.
- S. Chib. Estimation and comparison of multiple change-point models. *Journal of econometrics*, 86(2):221–241, 1998.
- S.B. Cho and H.J. Park. Efficient anomaly detection by modeling privilege flows using hidden Markov model. *Computers & Security*, 22(1):45–55, 2003.

- N. Chopin and F. Pelgrin. Bayesian inference and state number determination for hidden Markov models: an application to the information content of the yield curve about inflation. *Journal of Econometrics*, 123(2):327–344, 2004.
- S. Colella, C. Yau, J.M. Taylor, G. Mirza, H. Butler, P. Clouston, A.S. Bassett, A. Seller, C.C. Holmes, and J. Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, 35(6):2013, 2007.
- A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.
- R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.
- P.H.C. Eilers and R.X. De Menezes. Quantile smoothing of array CGH data. *Bioinformatics*, 21(7):1146–1153, 2005.
- P. Fearnhead and P. Clifford. On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899, 2003.
- J. Fridlyand, A.M. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90(1):132–153, 2004.
- X. Ge and P. Smyth. Deformable Markov model templates for time-series pattern matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 81–90. ACM, 2000.
- P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
- Y. Guédon. Exploring the state sequence space for hidden Markov and semi-Markov chains. *Computational Statistics & Data Analysis*, 51(5):2379–2409, 2007.
- S. Guha, Y. Li, and D. Neuberg. Bayesian hidden Markov modeling of array CGH data. *Journal of the American Statistical Association*, 103(482):485–497, 2008.
- J.A. Hartigan and M.A. Wong. Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.
- LI Hsu, S.G. Self, D. Grove, T. Randolph, K. Wang, J.J. Delrow, L. Loo, and P. Porter. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226, 2005.
- J.P. Hughes, P. Guttorp, and S.P. Charles. A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(1): 15–30, 1999.
- P. Hupé, N. Stransky, J.P. Thiery, F. Radvanyi, and E. Barillot. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422, 2004.

- D. Kimber and L. Wilcox. Acoustic segmentation for audio browsers. *Computing Science and Statistics*, pages 295–304, 1997.
- D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.
- W.R. Lai, M.D. Johnson, R. Kucherlapati, and P.J. Park. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21(19):3763, 2005.
- M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.
- Z. Liu, A. Li, V. Schulz, M. Chen, and D. Tuck. MixHMM: inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. *PloS one*, 5(6):e10909, 2010.
- T.M. Luong, Y. Rozenholc, and G. Nuel. Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model. *Arxiv preprint arXiv:1203.4394*, 2012.
- JC Marioni, NP Thorne, and S. Tavaré. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, 22(9):1144–1146, 2006.
- A.V. Nefian and M.H. Hayes III. Hidden Markov models for face recognition. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 5, pages 2721–2724. IEEE, 1998.
- A.B. Olshen, ES Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.
- F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(1):27, 2005.
- D. Pinkel, R. Seagraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20:207–211, 1998.
- L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- AE Raftery and VE Akman. Bayesian analysis of a Poisson process with a change-point. *Biometrika*, pages 85–89, 1986.
- G. Rigai, E. Lebarbier, and S. Robin. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, pages 1–13, 2011.
- O.M. Rueda and R. Díaz-Uriarte. Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Computational Biology*, 3(6):e122, 2007.
- G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.
- S.L. Scott. Bayesian methods for hidden Markov models. *Journal of the American Statistical Association*, 97(457):337–351, 2002.

- S.P. Shah, K. Cheung, N.A. Johnson, G. Alain, R.D. Gascoyne, D.E. Horsman, R.T. Ng, K.P. Murphy, et al. Model-based clustering of array CGH data. *Bioinformatics*, 25(12):i30, 2009.
- A.M. Snijders, N. Nowak, R. Segreaves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A.K. Hindle, B. Huey, K. Kimura, et al. Assembly of microarrays for genome-wide measurement of DNA copy number by CGH. *Nature Genetics*, 29:263–264, 2001.
- S. Stjernqvist, T. Rydén, M. Sköld, and J. Staaf. Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, 23(8):1006–1014, 2007.
- ES Venkatraman and A.B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663, 2007.
- A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.
- K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S.F.A. Grant, H. Hakonarson, and M. Bucan. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17(11):1665–1674, 2007.
- H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, 2005.
- N.R. Zhang and D.O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.
- Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on*, 20(1):45–57, 2001.

## A Sample R code

The following R code presents a toy example for data simulated from the Poisson distribution, with  $n = 100$  observations and  $J = 3$  different segments, with change-points after the observations 25 and 75. The previously described algorithms calculate the probability of an observation being in a hidden state (marginal) and being a change-point (cp).

The code for the *level-based* models:

```

n=100; # not usable for large n due to underflow issue
(logscale version needed for larger n)
L=3; # number of levels
lambda=c(4.5,3.0,7.0); # parameters for the three level
eta=0.03; # transition parameter
# transition matrix
pi=matrix(eta/(L-1),ncol=L,nrow=L);
diag(pi)=1.0-eta;

# generate the data according to the model
set.seed(42)
s=rep(NA,n); s[1]=1;
for (i in 2:n) s[i]=sample(1:L,size=1,prob=pi[s[i-1],]);
x=rpois(n,lambda[s]);
refcp=which(diff(s)!=0); # reference change-point location

# forward recursion
F=matrix(rep(NA,n*L),ncol=n);
F[,1]=0.0; F[1,1]=1.0*dpois(x[1],lambda[1]);
for (i in 2:n) for (l in 1:L)
  F[l,i]=sum(F[,i-1]*pi[,l]*dpois(x[i],lambda[l]));

# backward recursion
B=matrix(rep(NA,n*L),ncol=n); B[,n]=1.0;
for (i in seq(n,2,by=-1)) for (l in 1:L)
  B[l,i-1]=sum(pi[l,]*dpois(x[i],lambda)*B[,i]);

# probability of the evidence
pevidence=F[1,1]*B[1,1];

# consistency check TRUE if OK
# NOT RUN: prod(apply(F*B,2,sum)==pevidence)==1

# marginal distribution
marginal=F*B/pevidence;

# posterior distribution of change-points
cp=rep(0,n);
for (i in 2:n) for (l in 1:L)
  cp[i-1]=cp[i-1]+sum(F[l,i-1]*pi[l,-1]*dpois(x[i],lambda[-1])*B[-l,i])/pevidence;

par(mfrow=c(1,2));
plot(x,main="posterior marginal distribution of segment index",pch=16,col="blue");
abline(v=refcp,col="blue",lty=4); points(lambda[s],t="l",col="blue",lty=4);
for (j in 1:L) points(max(x)*marginal[j,],col=j,lty=j,t="l",lwd=2);
plot(x,main="posterior change-point distribution",pch=16,col="blue");
abline(v=refcp,col="blue",lty=4);
points(lambda[s],t="l",col="blue",lty=4);
points(max(x)*cp,t="l",lwd=2);

```

The code for the *segment-based* models:

```
n=100; # not usable for large n due to underflow issue
#(logscale version needed for larger n)
K=3; # number of segments
lambda=c(4.5,3.0,7.0); # parameters for the three segments
refcp=c(25,75); # ref change point locations
s=c(rep(1,25),rep(2,50),rep(3,25)); # true segment index

# generate the data according to the model
set.seed(42)
x=rpois(n,lambda[s]);

# forward recursion
F=matrix(rep(NA,n*K),ncol=n);
F[,1]=0.0; F[1,1]=1.0*dpois(x[1],lambda[1]);
for (i in 2:n) {
  F[2:K,i]=(0.5*F[1:(K-1),i-1]+0.5*F[2:K,i-1])*dpois(x[i],lambda[2:K]);
  F[1,i]=0.5*F[1,i-1]*dpois(x[i],lambda[1]);
};

# backward recursion
B=matrix(rep(NA,n*K),ncol=n); B[,n]=0.0; B[K,n]=1.0;
for (i in seq(n,2,by=-1)) {
  B[1:(K-1),i-1]=0.5*B[1:(K-1),i]*
    dpois(x[i],lambda[1:(K-1)])+0.5*B[2:K,i]*dpois(x[i],lambda[2:K]);
  B[K,i-1]=0.5*B[K,i]*dpois(x[i],lambda[K]);
};

# probability of the evidence
pevidence=F[1,1]*B[1,1];

# consistency check TRUE if OK
# NOT RUN: prod(apply(F*B,2,sum)==pevidence)==1

# marginal distribution
marginal=F*B/pevidence;

# posterior distribution of change-points
cp=matrix(0,nrow=K-1,ncol=n);
for (k in 1:(K-1)) cp[k,1:(n-1)]=F[k,1:(n-1)]/pevidence*0.5*B[k+1,2:n]*
  dpois(x[2:n],lambda=lambda[k+1]);

par(mfrow=c(1,2));
plot(x,main="posterior marginal distribution of segment index",pch=16,col="blue");
abline(v=refcp,col="blue",lty=4); points(lambda[s],t="l",col="blue",lty=4);
for (k in 1:K) points(max(x)*marginal[k,col=k],col=k,lty=k,t="l",lwd=2);
plot(x,main="posterior change-point distribution",pch=16,col="blue");
abline(v=refcp,col="blue",lty=4);
points(lambda[s],t="l",col="blue",lty=4);
for (k in 1:(K-1)) points(max(x)*cp[k,col=k],col=k,lty=k,t="l",lwd=2);
```

## B Sample datasets

The number of annual accidents from 1851 – 1962 in the British coal mining disaster data set [Carlin et al., 1992]:

```
accidents <- c(4, 5, 4, 1, 0, 4, 3, 4, 0, 6, 3, 3, 4, 0, 2, 6, 3, 3, 5,  
4, 5, 3, 1, 4, 4, 1, 5, 5, 3, 4, 2, 5, 2, 2, 3, 4, 2, 1, 3, 2, 2, 1, 1,  
1, 1, 3, 0, 0, 1, 0, 1, 1, 0, 0, 3, 1, 0, 3, 2, 2, 0, 1, 1, 1, 0, 1, 0,  
1, 0, 0, 0, 2, 1, 0, 0, 0, 1, 1, 0, 2, 3, 3, 1, 1, 2, 1, 1, 1, 1, 2, 4,  
2, 0, 0, 0, 1, 4, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1)
```

The log-reference ratio (LRR) of chromosome 10 of cell line BT474 in breast cancer tumor [Snijders et al., 2001]:

```
LRR <- c(0.3362, 0.2793, 0.3742, 0.4424, 0.1238, 0.3590, 0.1655, 0.3552,  
0.1504, 0.2983, 0.3173, 0.1428, 0.1276, 0.8824, 0.3438, 0.2642, 0.1390,  
0.3211, 0.4235, 0.2338, 0.2983, 0.2376, 0.2452, 0.3362, 0.3400, 0.3248,  
0.3666, 0.4766, 0.7193, 0.3135, 0.1883, 0.1466, 0.3211, 0.3779, 0.2376,  
0.1655, 0.3173, 0.0252, 0.2528, 0.3324, 0.2528, 0.2755, 0.2945, 0.1997,  
0.2376, 0.1807, 0.6510, 0.3552, 0.1883, 0.1655, 0.2680, 0.2642, 0.2680,  
0.1883, 0.1997, 0.1693, 0.3362, 0.2111, 0.2755, 0.2680, 0.2680, 0.2983,  
0.1162, 0.3476, 0.3817, 0.8331, 0.3097, 0.2376, 0.0556, 0.0707, 0.1655,  
0.1769, 0.2111, -0.0696, 0.2149, 0.0214, 0.1238, -0.7333, -0.7409,  
0.0366, 0.0897, 0.1769, 0.0404, 0.0935, 0.1466, 0.6169, 0.1921, 0.2300,  
0.3476, 0.1921, 0.6055, 0.0935, 0.0518, -0.0582, -0.6423, 0.4083,  
-0.7864, -0.8964, -0.6120, -0.7258, -0.6347, -0.7940, -0.6120, -0.6006,  
-0.1986, -0.6044, -0.4754, -0.6234, -0.7030, -0.5323, 0.3097, -0.8395,  
-1.0064, -0.8206, -0.8851, -0.0506, -0.7409, -0.7296, -1.5526, -0.1455)
```